Tõenäosusteooria ja statistika II

Kodutöö
Näidislahendus

Ülesanne 1 – hinnangu täpsuse kirjeldamine

Sissejuhatus:

Soovime kogutud andmete põhjal (elementaarosakeste kokkupõrgete arvud ajaühiku jooksul) hinnata tõenäosust, et saame uuritava tunnuse väärtuseks nulli. Kogutud andmed ise on järgmised:

2 3 2 3 3 2 1 0 3 0 2 0 1 0 1 0 1 1 3 1 1 0 2 0 0

Nulli saamise tõenäosust võime hinnata kahel viisil:

Meetod 1: $\hat{p}_0 = \frac{nullide\ arv}{vaatluste\ arv}$ ehk võime kasutada nulli nägemise suhtelist sagedust.

Meetod 2: Oletame, et uuritav tunnus on Poissoni jaotusega (kokkupõrgete arv võiks käituda kui Poissoni jaotusega juhuslik suurus). Hindame Poissoni jaotuse parameetri (suurima tõepära hinnang Poissoni jaotuse parameetrile on kõigi vaatluste keskmine, $\hat{\lambda} = \overline{X}$). Leiame, milline on Poissoni jaotusega juhusliku suuruse (parameetriga $\hat{\lambda}$ Poissoni jaotus) tõenäosus omandada väärtust 0 ja kasutame saadud tõenäosust kui oma hinnangut.

Küsimused:

a) Hinda nulli nägemise tõenäosust mõlema kirjeldatud meetodi abil.

Vastus:

Meetod 1: $\hat{p}_0 = \frac{8}{25} = 0{,}32$

Meetod 2: $\hat{\lambda} = 1{,}28$; P(X=0) = $\hat{\lambda}^0 \exp(-\hat{\lambda}) / 0! = \exp(-\hat{\lambda}) \approx 0{,}278$

b) Lisa mõlemale hinnangule ka hinnangu standardviga.

Vastus:

Meetod 1 – hinnangu standardvea hinnang. Tähistame Z indikaatortunnust, mis näitab, kas vaatlus on 0 või mitte (Z=1, kui vaatlus on 0 ja Z=0 muude vaadeldud väärtuste korral). Sellisel juhul $\overline{Z} = \hat{p}_0$ ja $\widehat{se_0} = \sqrt{\frac{D(Z)}{n}} = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}} \approx 0{,}093$.

Meetod 2
Leiame esmalt standardvea delta meetodit kasutades.
Teame, et $\overline{X} \sim N\left(\mu = \lambda; \sigma^2 = \frac{\lambda}{n}\right)$, sest Poissoni jaotuse dispersioon on võrdne keskväärtuse ehk Poissoni jaotuse parameetriga λ. Delta meetod ütleb, et sellisel juhul, ligikaudu, $g(\overline{X}) \sim N\left(\mu =\right.$

$g(\lambda); g'(\lambda)^2 \frac{\lambda}{n}\Big)$ ehk transformeeritud hinnangu $g(\overline{X})$ dispersioon on $g'(\lambda)^2 \frac{\lambda}{n}$. Antud ülesandes on g(x)=exp(-x) ja seega tuleb hinnagu standardvea hinnanguks:

$$\sqrt{\left(\exp\left(-\hat{\lambda}\right)(-1)\right)^2 \frac{\hat{\lambda}}{n}} = \sqrt{\frac{\exp(-1{,}28)^2 \cdot 1{,}28}{25}} \approx 0{,}0629$$

Alternatiivina võime kasutada parameetrilist bootstrap-meetodit standardvea hindamiseks:

```
# Parameetriline bootstrap
lambda=mean(x)
bhinnang=rep(NA, 10000)
for (i in 1:10000){
  bvalim=rpois(25, lambda)
  bhinnang[i]=exp(-mean(bvalim))
}
sd(bhinnang)
```

Mis annab meile praegu standardvea hinnanguliseks väärtuseks 0,0636. Parameetrilist bootstrap-meetodit eelistame mitteparameetrilisele bootstrapile praegu sellepärast, et tegemist on väikese valimiga.

c) Põhjenda kumba meetodit kasutaksid sina nulli esinemistõenäosuse hindamisel ja miks eelistad just seda meetodit.

Vastus:
Teine meetod suudab nulli nägemise tõenäosust täpsemalt hinnata – seega eelistaksin teist meetodit (muidugi juhul, kui saaksime olla täiesti kindlad, et uuritava tunnuse tegelikuks jaotuseks on ikka tõepoolest Poissoni jaotus).

Ülesanne 2 - Suurima tõepära hinnangu leidmine numbrilise maksimiseerimise abil

Ühe küsitluse käigus mõõdeti tudengite pikkuseid. Mõõtmistulemused saad R-i lugeda järgmise käsu abil:

```
print(load(url("http://www-1.ms.ut.ee/mart/TS2/pikkused.RData")))
hist(pikkused)
# Esimese kümne tudengi pikkused:
pikkused[1:10]
```

Soovime hinnata nii nais- kui ka meestudengite keskmist pikkust; nii naiste kui ka meeste pikkuste varieeruvust (standardhälbeid) ja muidugi ka naiste osakaalu. Paraku pole meil kirjas mis soost tudeng ühe või teise pikkusega oli – meil on ununenud mõõtmata tudengi sugu.

Õnneks teame üsna kindalt, et nii meeste kui ka naiste pikkuste jaotuseks võiks olla normaaljaotus – pikkuse tinglikud jaotused tingimusel et sugu on antud on normaaljaotused. Seega on pikkuste marginaaljaotus kirja pandav kui jaotuste segu (meenuta täistõenäosuse valemit):

$$F_{PIKKUS}(x) = naiste\_osakaal\ F_{PIKKUS|Naine}(x) + (1-naiste\_osakaal)\ F_{PIKKUS|Mees}(x)$$

Ehk, alternatiivselt kirja pandult:

$$f_{PIKKUS}(x) = naiste\_osakaal\ f_{PIKKUS|Naine}(x) + (1-naiste\_osakaal)\ f_{PIKKUS|Mees}(x)$$

kus $f_{PIKKUS|Naine}(x)$ on parameetritega μ_{naine}; σ_{naine} normaaljatuse tihedusfunktsioon ja $f_{PIKKUS|Mees}(x)$ on parameetritega μ_{mees}; σ_{mees} normaaljatuse tihedusfunktsioon.

Küsimused:

1. **Milline näeb välja log-tõepära tudengi pikkuste andmete jaoks? Pane kirja R-i funktsioon mis leiab etteantud parameetrite vektori jaoks log-tõepära väärtuse.** Antud juhul on tudengite pikkuste jaotusel viis tundmatut parameetrit: μ_{naine}; σ_{naine}; μ_{mees}; σ_{mees} ja naiste_osakaal. ,

```
l=function(arg, andmed){
   mu1=arg[1]
   sigma1=exp(arg[2])

   mu2=arg[3]
   sigma2=exp(arg[4])

   # expit
   osakaal=exp(arg[5])/(1+exp(arg[5]))

   l=sum(log( osakaal*dnorm(andmed, mean=mu1, sd=sigma1)+
          (1-osakaal)*dnorm(andmed, mean=mu2, sd=sigma2) ) )
   return(l)
}
```

Antud programmis argument arg on hinnatavate parameetrite vektor, kusjuures esimese grupi osakaal on leitav vektori arg 5. elemendi järgi järgmiselt: osakaal=exp(arg[5])/(1+exp(arg[5])) ehk osakaal=expit(arg[5]). Sellist transformatsiooni kasutame seetõttu, et esimese grupi (naiste) osakaal peab alati jääma vahemikku 0…1. Nimelt on sõltumata arg[5] väärtusest taolisel viisil transformeeritud väärtus alati vahemikus 0…1.

2. **Hinda nende viie parameetri väärtused numbriliste meetodite (näiteks optim-käsu) abil**. Muretse ka selle pärast, et kõigi parameetrite hinnangud jääksid lubatud piiridesse (näiteks ei soovi me näha negatiivseid standardhälbeid). Millised hinnangud saad? Kommenteeri saadud hinnanguid, on need usutavad? Kui mitte, siis arutle selle üle, mis võis valesti minna.

   Vastus:
   Naiste keskmine pikkus on hinnanguliselt 168,3 cm; naiste pikkuste standardhälve on 5,9. Meeste keskmine pikkus on hinnanguliselt 182,7cm ja meeste pikkuste standardhälve on 6,9. Hinnanguliselt 80,3% uuritavasse populatsiooni kuuluvatest inimestest naised. Naiste keskmine pikkus tuli meeste pikkusest väiksem, samuti on hinnang naiste pikkuste standardhälbele veidi väiksem kui hinnang meeste pikkuste standardhälbele. Kirjeldatud hinnangute järgi võiks 95% naiste pikkused jääda vahemikku 156,6cm … 180,1cm, mis on usutav. Meeste keskmine pikkus +-2 standardhälvet annab ootuspärase vahemiku meeste pikkustele: 168,9cm..196,5cm. Antud vahemik paistab samuti hästi kirjeldavat tüüpilist meest. Seega on leitud hinnangud usutavad.

3. **Lisa tööle kasutatud R-i programm** (milline näeb välja log-tõepära väärtust arvutav programm; kuidas kasutasid optim-käsku; kuidas optim-käsu tulemustest lugesid välja hinnangute väärtused…

```
print(load(url("http://www-1.ms.ut.ee/mart/TS2/pikkused.RData")))

l=function(arg, andmed){
 mu1=arg[1]
 sigma1=exp(arg[2])

 mu2=arg[3]
 sigma2=exp(arg[4])

 # expit
 osakaal=exp(arg[5])/(1+exp(arg[5]))

 l=sum(log(   osakaal*dnorm(andmed, mean=mu1, sd=sigma1 )+
      (1-osakaal)*dnorm(andmed, mean=mu2, sd=sigma2 )   ))
 return(l)
}

tul=optim(c(160,2, 180, 2, -0.5), l, andmed=pikkused,
      control=list(fnscale=-1))
tul=optim(tul$par, l, andmed=pikkused,
      control=list(fnscale=-1))
tul=optim(tul$par, l, andmed=pikkused,
      control=list(fnscale=-1))

# Hinnatud keskväärtus
tul$par[1]
# Hinnatud standardhälve
exp(tul$par[2])

# Hinnatud keskväärtus
tul$par[3]
# Hinnatud standardhälve
exp(tul$par[4])

# Naiste osakaal
n_osakaal = exp(tul$par[5])/(1+exp(tul$par[5]))
n_osakaal
```